Algorithms

- Algorithms for LPs
  - Naive: enumerate all bases
  - Simplex
  - Constraint/variable generation (sort of)

- Algorithms for QPs
  - Constraint generation

- Upcoming:
  - Subgradient descent
  - Ellipsoid
  - Interior point (later)
What’s a subgradient?

\[ f(y) \geq f(x) + (y-x) \cdot g \quad \forall y \]

\[ \iff g \in \partial f(x) \]
Residents of Delphi

subgradient

separation oracle for epigraph

$h(x)$

$(x, v) \quad v < f(x)$
Working w/ subgradients

• Subgradients are just like gradients (sort of):
  ‣ when $f'(x)$ exists: $\partial f(x) = \{ f'(x) \}$
  ‣ $\partial (f+g)(x) = \partial f(x) + \partial g(x)$
  ‣ $\partial f(Ax+b) = A^T(\partial f)(Ax+b)$
  ‣ $\partial g(x) = \partial (f(g(x))) = \partial f(g(x)) \partial g(x)$
  ‣ $0 \in \partial f(x) \iff x \in \text{arg min } f(x)$

$f, g$ convex
Proof of chain rule

\[ x_0 \in \partial g(x_0) \quad y_0 = g(x_0) \quad v \in \partial f(y_0) \]

• \( \partial f(g(x)) = (\partial f)(g(x)) \partial g(x) \)

  • \( f, g \) convex; \( f \) monotone

  \[ g(x) \geq g(x_0) + (x - x_0) \cdot u \]

  \[ f(g(x)) \geq f(g(x_0)) + (x - x_0) \cdot u \cdot \]

  \[ f(y) \geq f(y_0) + (y - y_0) \cdot v \]

  \[ f(g(x)) \geq f(g(x_0)) + (x - x_0) \cdot u \cdot \]

  \[ f(g(x)) \geq f(g(x_0)) + (x - x_0) \cdot u \cdot \]

  \[ \iff \quad u \in \partial f(g(x)) \]
Ex: dual norm

- Given a norm, \( \|x\| \)
- The **dual norm** is
  - \( \|y\|_* = \max_{x \in C} x^T y \)
  - \( C = \{ x \mid \|x\| \leq 1 \} \)
  - ex: dual of \( \|x\|_1 \) is: \( \|y\|_{\infty} = \max_{i} |y_i| \)
  - ex: dual of \( \|x\|_2 \) is: \( \|y\|_2 \)

- Suppose we want to solve
  - \( \min_w \|w\|_*^2 + \lambda \sum_i (x_i^T w - y_i)^2 \)
Subgradient of dual norm

- \( \|w\|_* = \max_{x \in C} x^T w \) where \( C = \{ x \mid \|x\| \leq 1 \} \)
- \( \partial \|w\|_*^2 = 2 \|w\|_*^2 \|w\|_* \)
- Let \( x_w \in \arg \max_{x \in C} x^T w \)
  - \( \|w\|_* = w \cdot x_w \)
  - \( \|v\|_* \geq v \cdot x_w \)
  - \( \|u\|_* - \|w\|_* \geq (u - w) \cdot x_w \)
- \( \partial \|w\|_* = \{ \arg \max_{x \in C} x \cdot w \} \)
Ex: SVM subgradient

- \(\min_w L(w) = \|w\|^2/2 + (C/m) \sum_i h(y_i(x_i^T w - b))\)
  - where \(h(z) = \max(0, 1 - z)\)

- \(\partial h(z) = \begin{cases} 
0 & z \geq 1 \\
(z - 1) & -1 \leq z < 1 \\
-1 & z < -1
\end{cases}\)

- \(\partial L(w) = w + \sum_{i} \alpha_i y_i x_i\)
SVM loss: hinge portion
SVM loss: hinge portion
SVM loss: $C = 10$
Subgradient descent

- Greedily try to decrease objective
  - fastest local decrease: negative subgradient if smooth
Subgradient descent

• Initialize $x_1$
  
• for $t = 1$ to “I’m tired”
  
    ‣ $g_t = \text{any element of } \partial f(x_t)$
    
    ‣ $x_{t+1} = x_t - \eta_t g_t$

• Questions:
  
    ‣ how to initialize?
    
    ‣ when are we tired?
    
    ‣ what to use for $\eta_t$?
Numerical example

• SVM, training examples:
  ▸ $(3, +), (0.1, –)$

• $L(w, b) = \frac{w^2}{2} + \frac{C}{2}(h(z_1)+h(z_2))$
  ▸ $z_1 = 3w - b$
  ▸ $z_2 = -0.1w + b$

• $\partial L(w, b) = \begin{pmatrix}
  \frac{C}{2} (3\alpha_1 - \alpha_2) \\
  \frac{C}{2} (-\alpha_1 + \alpha_2)
\end{pmatrix}$
Subgradient path
Subgradient projection

- Initialize $x_1$
- for $t = 1$ to “I’m tired”:
  - $g_t =$ any element of $\partial f(x_t)$
  - $x_{t+0.5} = x_t - \eta_t g_t$
  - $x_{t+1} = \text{arg min}_{x \in F} ||x - x_{t+0.5}||$

Problem:

$$\min_x f(x) \quad \text{s.t.} \quad x \in F$$

Example of why subgradient projection is different from just projecting the unconstrained minimum onto the constraints: minimize elliptical quadratic s.t. L1 constraint

Example: max $x+y$ within unit circle
Stochastic subgradient

• Recall SVM subgradient:
  \[ \partial L = w + (C/m) \sum_{i=1}^{m} y_i x_i \partial h(y_i(x_i^T w - b)) \]

• If many examples: select \( m_0 \ll m \) examples at random, compute
  \[ \hat{\partial L} \approx w + (C/m_0) \sum_{i \in S} y_i x_i \partial h(y_i(x_i^T w - b)) \]
  where \( S \) is a selected set

• Simple trick: to get an unbiased estimate of \( \partial L \), select \( m_0 \ll m \) examples at random, compute
  \[ \partial L = w + (C/m_0) \sum_{i \in S} y_i x_i \partial h(y_i(x_i^T w - b)) \]

\[ L \approx \|w\|^2/2 + (C/m_0) \sum_{i \in S} y_i x_i h(y_i(x_i^T w - b)) \]
Example: $m_0 = 10$ of 100
Stochastic subgradient (or stoch. subg. projection)

• Initialize $x_1$

• for $t = 1$ to “I’m tired”:
  ‣ $f_t =$ estimate of $f$
  ‣ $g_t =$ any element of $\partial f_t(x_t)$
  ‣ $x_{t+0.5} = x_t - \eta_t g_t$
  ‣ $x_{t+1} = \text{arg min}_{x \in F} ||x-x_{t+0.5}||_Z$ (or just $x_{t+0.5}$)

Problem:
\[
\min_x f(x)
\] (optionally, s.t. $x \in F$)
Strict convexity

- Def’n: \( f \) is \( \lambda \)-strictly convex if, for \( g \in \partial f(x) \),
  \[ f(y) \geq f(x) + (y - x) \cdot g + \lambda \|y - x\|_2^2 / 2 \]

- note: \( \lambda = 0 \) \( \iff \) ordinary convexity
A useful inequality

\[ \sum_{i=1}^{t} \frac{1}{x_i} \leq 1 + \ln t \]

\[ 1 + \int_{1}^{t} \frac{1}{x} \, dx \]

\[ 1 = [\alpha x]_1 \]

\[ 1 + \ln t - \theta \]
Convergence preview

• For strictly convex $f(x)$ (i.e., $\lambda>0$):
  
  ‣ set $\eta_t = \frac{1}{\lambda t}$
  ‣ $f(x_t) - f(x^*) = O\left(\frac{1}{t}\right)$

• For non-strictly convex $f(x)$ (i.e., $\lambda=0$):
  
  ‣ set $\eta_t = o\left(\frac{1}{\sqrt{t}}\right)$
  ‣ $f(x_t) - f(x^*) = o\left(\frac{1}{\sqrt{t}}\right)$

• To get accuracy $\varepsilon$:
  
  ‣ $\lambda>0$: $T = O\left(\frac{1}{\varepsilon}\right)$
  ‣ $\lambda=0$: $T = o\left(\frac{1}{\varepsilon^2}\right)$

Interior point:
$T = O\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$
but each iter slower